Review of the manuscript “Implementation of a resolved building capability in the Energy Research and Forecasting model for accelerated high-fidelity urban-scale simulation.”

General comments:

The paper presents and validates a newly implemented resolved building feature added to the exascale Energy Research and Forecasting (ERF) model, set up to predict wind and particle dispersion during the Joint Urban 2003 Oklahoma City tracer release experiment. The model examines scenarios under both stable and unstable thermodynamic conditions, focusing on building wall roughness length as a key parameter for near-wall parameterization. Various model configurations are tested, including adaptive mesh refinement and both compressible and anelastic solvers.

The content is relevant to the scope of GMD, and the results merit publication; however, the manuscript needs revisions in several areas because the material is not always sufficiently detailed or accurate. Below are specific comments on aspects that require improvement.

Specific comments:

In the abstract, introduction, and other sections, it is stated that the model effectively addresses critical factors for providing accurate forecasts on exascale computing platforms to guide officials, ensure public safety, or support emergency response. Several points in this context need clearer explanation.

1. It is not clear whether the ERF model is currently fulfilling any official duties, or this is only the research development. Please clarify this in the appropriate sections where the application for public safety/emergency is mentioned.

2. The ERF model is applied for the convective and stably stratified flows, however the current wall parameterization does not account for the buoyancy effects which is clearly the drawback, and this development is planned in the future. It would be helpful to elaborate on the significance of this missing feature and how it might affect the presented results, particularly if the model is currently used or intended for any official applications.

Model’s performance.

The presented simulations were using up to 128 AMD MI300A GPU accelerated processing units on LLNL’s Tuolumne HPC, with the scalability results presented in section 2.1. From the official information, the whole LANL system has 288 Petaflop peak performance combined from 4608 GPU units and 110592 CUP cores, so in reality this is not the exascale system. The single peak performance of the CDNA 3 AMD MI300A GPU is 61 TFlops for the double precision computations. For 128 GPU units this sums in 7.8 PFlops, thus the current experiments are at least two orders of magnitude below the exascale in terms of the peak performance. Moreover, sustained performance of real CFD applications on GPU systems is usually at the level of 2-3% of peak performance.

3. Please calculate the actual sustained performance of the ERF model using the tested architecture and update this information, so that readers can gain a clearer understanding of your model's position on the exascale roadmap. You may add the scalability plot in section 2.1 with the real sustained performance in Flops units.

4. Introduction: “in section 5, we compare different model configurations (e.g., compressible or anelastic mode) to explore accuracy and performance tradeoffs” and further at the end of section 2.1: “The increase in performance for using anelastic mode compared to fully compressible mode will vary and need to be optimized from problem to problem; however, significant performance gains should be expected. In later sections, we explore the performance versus accuracy tradeoffs of the different modes”. Unfortunately, it seems the only discussion relevant to model performance is in section 2.1, a general discussion is also presented on model efficacy (a tradeoff between performance and accuracy).

   Additional more detailed analysis is necessary for two key components. What is the computational performance or overhead of the option with wall scheme comparing to the original version. And what is the overhead of the AMR option, which provides additional high resolution subdomain which area is growing in time according to the area covered by the real plume spread.

The literature overview.

5. Introduction l16-17 “Accurate simulation is required to provide guidance to officials to ensure public safety; however, existing high-fidelity urban modeling tools are too slow to be used for time-sensitive applications” … please provide examples. As far as I know there are plenty of different tools and models which are relatively efficient and used for decades in the high-fidelity/fast response modeling supporting the public safety needs, for example: QUIC-URB, HPDM, CFD-Urban, ADMS, etc. How much is the current system improving this state?

6. Introduction l18-20 ”buildings have highly complex geometries and properly resolving these structures requires very small grid spacing, typically on the scale of a few meters, which is challenging using large-eddy simulation (LES) codes developed to run on traditional HPC” …. while this is true, for decades the standard approach was to use RANS models which are computationally more efficient than LES. Authors should mention RANS methodology and discuss their pros and cons in relation to LES approach implemented in ERF model.

7. Authors state at a couple of instances (Introduction l57-58 and section 2.2. l135) that immersed body forces in relation to urban modeling “was first investigated by Chan and Leach (2007) who demonstrated that for larger drag terms, the flow around a cube acted similarly to when modeling the cube as a solid boundary”. The FEM3MP model used by referenced work of Chan and Leach 2007 did not use the LES approach but the simplified liner-eddy-viscosity (LEV) model which had no explicit treatment for building-induced turbulence. Otherwise, the more advanced version nonlinear-eddy-visosity (NEV) did not account for nonneutral stability conditions. There were different LES models applied at similar time to urban modeling which use immersed body forces, as e.g. EULAG model from Smolarkiewicz et al. (JCP, 227(1), 663-653, 2007) applied for Pentagon studies, which was further validated over the JU2003 IOP’s in Wyszogrodzki et al. (AR, 118, 324-345, 2012). That model accounts for both the building-generated turbulence and the stability effects, so it may be used as an example of the first development in this respect. Pleae, add these references and clarify/update your statements.

More statements to be corrected or clarified.

8. Introduction: l22-23: “Moreover, these (accelerated high fidelity) tools enable running ensembles of simulations and performing sensitivity studies to provide more robust guidance and to inform lower-order faster running models”. I’m not sure what is the reason for using low-order faster models on top of the ensemble of GPU accelerated high-fidelity ensemble which is on its own able to produce fast and accurate solution? Do you refer to the so called PIML (Physically Informed Machine Learning) approach … if so, describe more clearly this idea.

9. Introduction l29-32: “The AMR capabilities of ERF enable further performance gains by simulating only certain, necessary regions with higher resolution while simultaneously simulating other regions at coarser grid spacing” … please add here or in further sections more details on AMR configuration, whether this a one- or two-way nesting, which variables are prolongated, this especially concerns the turbulence component. How the boundary conditions are set up for the denser grid for both compressible and anelastic option, which may not be so straightforward as the model uses Arakawa C-Grid formulation. Which prolongation and reduction algorithms are applied, are they first or higher order accurate?

10. Section 2.1. Authors refer to a couple of options within ERF to be exactly the same or similar as in WRF model (grid nesting, spatial discretization, advection RK schemes, terrain-following coordinate and stretching) … so is the ERF model being developed as a simplified version of WRF model or it is completely different application build on the same as WRF concept? Maybe it would be wiser to put an effort and extend WRF-IMB version of Lundquist et al. (2010, 2012)?

11. Section 2.1. “The anelastic method used here employs a fast Fourier transform based solver for the perturbational pressure (but a multigrid solver is also available)” … please add more details on the anelastic solver or references to such details, add also information on how fast is the convergence of the elliptic solver, is it dependent on the grid/nesting configuration?

12. Section 2.1 “we run a simple scaling test similar to that conducted by Sauer and Munoz-Esparza 110 (2020) … Buildings are also included in the simulations (we include a subset of buildings from Torino, Italy) to ensure that the performance demonstrated is what a user should expect when running building-resolving simulations in ERF” …. why is Torino/Italy mentioned in this place? The work of Sauer and Munoz-Esperaza is not applicable for urban modeling, why is this work used as the basis for performance expectation in building resolving application?

13. Section 2.2 in presenting the wall model. It seems there is completely no difference between walls and rooftops as the log formulations e.g. 1-3 are identical, and the only difference is in tangential velocity components. The current ERF model accounts for the thermal stratification, however the wall model does not include the thermal effects acting through the buoyancy force. However, even in this case there should be an effect of the gravity accounted to differentiate between walls and the rooftop. So, the question is what the boundary conditions for the pressure at the rooftop are, since pressure gradients act over there as the balance to the gravitational force. Should these conditions be different for the fully compressible and anelastic formulations, where the later applies either the fast Fourier or the multigrid solvers.

If no special pressure condition is applied, then the residual pressure may still be present inside the solid building structures affecting the flow outside. Please discuss this topic in this section.

14. Section 5.1.1. “While buildings come in many forms and a single roughness value cannot be representative of all buildings, we simply do an order of magnitude analysis to determine the approximate order of magnitude that the roughness value should be” …. I’m not sure if selection of a single roughness value for all is the best option to determine the proper expected magnitude. Maybe better approach would be to differentiate on the basis of the building size or type as the residential areas, the downtown skyscrapers, etc.?

15. Section 5.1.1 “As the flow evolves and the SF6 concentration disperses into a clear plume, the AMR grid continually expands downwind and vertically refining the required region” … how is the AMR controlled? … the nested grid is not extending over the plume area, as the plume is dispersing and turning west, while nest is centered and not extending to the lateral boundary.

16. Section 5.1.1. “As the building wall roughness length increases the near-surface wind speed decreases, which is most evident in the slower wind speeds downwind of the central business district and near the smaller residential buildings to the north. These slower wind speeds are due to the increased drag imparted on the flow with the larger roughness length. Qualitatively, the general behavior of the wind speeds in the central business and flow speedup in the street canyons is well-predicted by model regardless of building wall roughness length” … the last sentence seems to be contradicting with the previous which describes different flow features (as wind speed) dependent on the roughness length. Does it mean that in central business district and in the street canyons model behaves always qualitatively well regardless of the wall model and roughness length used? Is then the wall model necessary at all?

17. Section 5.1.1. “Figure 11. While the larger drag imparted on the flow does increase mixing in general, the wind speeds at street level and in the central business district are very small, smaller than when using smaller building wall roughness lengths, thus reducing dispersion” … change terms small/smaller with specific numeric values.

18. Section 5.1.1 “Overall, 81% of stations are within a factor of two for the z0,w = 0.1 m case with 79% for the z0,w = 0.01 m case (there are 68 total station in the simulated area). For z0,w = 0.01 m case, 97% of stations are within a factor of five while 100% of stations are within a factor of five for the z0,w = 0.1 m case. For the z0,w = 1.0 m case, only 93% of stations are within a factor of five resulting in the lowest model skill of the evaluated configurations (see Table 1)” …. This information looks a bit messy, needs to be written in a clearer way, also add how many stations were within the factor 2 in z0,w=1.0 case.

19. Section 5.1.1. “Ultimately, we conclude that the z0,w = 0.1 m case performs the best. While the FB and NMSE for the z0,w = 0.1 m case are slightly less accurate compared to the z0,w = 0.01 m case, a larger number of stations are within both a factor of two and a factor of five, which the authors weigh more heavily” … FAC2/FAC5 statistics presented in Table 1 show that for z=0.1 there is more stations within FAC2 and more stations within FAC5 than for z=0.01, shouldn’t FB/NMS statistics be consistent, what the “heavy weighting” means in this context?

20. Section 5.1.2 Why there is no anelastic version at the 2.5 resolution. Earlier statements indicate that anelastic version is more efficient, so maybe it should be added for comparison?

21. Section 5.1.2 The discussion concerns only the resolved TKE, why the subgrid part is not accounted?

22. On submeso motions: Section 5.2 “there are a number of stations upwind and to the west of the tracer release location that are significantly underestimated. This underestimation is likely because of missing submeso motions as has been similarly discussed by Wiersema et al. (2020) and Wiersema et al. (2022). An idealized setup is limited in capturing the larger-scale turbulent structures that multiscale (mesoscale to microscale) setups are able to capture” and in Conclusions “measurement stations on the outer edges of the central business district that were underestimated likely due to missing submeso turbulent structures important for meandering of the plume” …. what authors mean in this case as for the submeso eddy or sumeso turbulent structure? Is that a single convective cell within boundary layer or smaller coherent structures relevant to turbulence fluctuations? Predicting the exact realistic convective structures which are responsible for plume meandering at short time scales is impossible for the current NWP models working at resolutions above LES (> 150 m). Introducing direct eddy structures generated by NWP model running at crude resolution may be an additional source of error. For the plume meandering, variable in time wind profile (wind speed, direction, shear) should be efficient to generate realistic plume meandering.

23. Conclusion” “Using a more realistic, multi-scale setup including land surface variability, terrain, and mesoscale parameterizations should improve model skill and is the subject of ongoing work” … which mesoscale parameterizations authors have in mind?

24. Conclusions L560, “ The prediction of SF6 concentration for the convective case study resulted in slightly larger errors or less model skill compared to the stable case” … be more specific in exact numbers or percentages.

25. Appendix “the flow behavior between the Δ = 10 mm and Δ = 20 mm cases indicating that it is preferable to have at least 20 grid cells spanning a building” … I’m not sure whether number of cells spanning the building is the correct parameter which determines an ability of the model for resolving the flow behind the building. You can repeat nesting experiments, in one you can refine only the area to left from recirculation zone to have more grid points over the building and the same number in recirculation zone, and in second case add nesting only covering the recirculation zone outside the building. Will these modifications reproduce closed recirculation as in previous Δ = 10 mm case or open recirculation as in Δ = 20 mm case?

Minor editorial issues:

Introduction l28: “is one such code and is built on the AMReX framework” … you can delete “and is”

Figure 1 caption “on AMD AMD MI300A” remove one of AMD’s … and “Tuolumne HPC machine” you may substitute “machine” with cluster, supercomputer or computing system

L 168 “method to determine u_target from one of two ways” … not clear what do you mean by “two ways”, maybe change to “method to determine u_target depending, whether cells are representing solid or the boundary”

L224 “higher heights” … maybe “higher levels” or “greater heights”?

L286 “compared to having no heat flux” … change to e.g. “comparing to no flux option”

L331: “Towards the beginning” … maybe “at the beginning” or “after start” vs “towards the end”

Figure 7. … “The red dashed line in (b) corresponds to the vertical cross sections shown in Figure 8 are taken from” ….. this sounds not correct from the grammatical point of view.

Figures 9 and 17 looks very impressive from the visual point of view but they are not introducing any new information over the crossection presented at fig 7, 8, 15, 16. To save space and consistency I suggest moving these graphics to supplemental material where you can add more graphics and even animations which present the three dimensionality of these results.

L564 “minimum threshold for inclusion for comparison” … change to e.g. “minimum threshold included for comparison”

Appednix A

L603 “sensitive to a number of other input parameters. Namely, these parameters include the turbulence closure” …. these are rather not “input parameters” but model configurations, algorithms or parametrizations

L604 „ …(Smagorinsky vs. TKE-1.5). the solver “ … delete dot

Bibliography:

L698 “IMMERSED BOUNDARY METHODS” …. change uppercase to lowercase